УДК 519.17

Москалец Р. Ю., Блеканов И. С.

Теоретико-графовые характеристики в вебометрических исследованиях внутренней топологии крупных сегментов Веба

1. Введение. В настоящее время стремительно развивается молодое научное направление вебометрика [1–3], которая занимается исследованиями количественных и качественных характеристик топологии (гиперссылочной структуры) различных веб-сегментов. Основоположниками таких исследований являются испанская группа Cybermetrics Lab, которая разработала вебометрический рейтинг сайтов различных крупных организаций [4], таких как научно-образовательные учреждения, больницы, научно-исследовательские центры и т. п.

Одной из актуальных задач вебометрики является задача анализа внутренней топологии различных сайтов [2, 5–7], а также выявление критериев оценки качества веб-сегментов и сравнение этих сегментов по полученным оценкам.

Для решения вышеуказанной задачи в статье рассмотрены известные теоретико-графовые характеристики веб-графов и разработан комплекс программ на их основе. Данный программный комплекс используется для вычисления указанных выше характеристик, их сравнения на разных веб-сегментах большой размерности, а также для визуализации результатов сравнительного анализа.

2. Эксперимент. В работе ставился эксперимент, в котором разработанный авторами комплекс программ для выявления значимых страниц внутренней топологии крупных веб-сегментов (с помощью используемых теоретико-графовых характеристик) и их визуализации апробировался на крупных сайтах университетского Веба.

Получение внутренней топологии веб-ресурсов в эксперименте выполнялось с помощью программно-аналитического комплекса для

Москалец Роман Юрьевич – студент, Санкт-Петербургский государственный университет; e-mail: rmoskalets@gmail.com, тел.: +7(951)672-76-91

Блеканов Иван Станиславович – доцент, Санкт-Петербургский государственный университет; e-mail: i.blekanov@gmail.com, тел.: +7(921)339-53-43

Работа выполнена при финансовой поддержке РФФИ, грант № 15-01-06105

вебометрических исследований, основанного на обобщенном ядре поискового робота [8] и успешно апробированного в исследованиях [5–7].

В качестве теоретико-графовых характеристик сайтов были выбраны несколько мер центральности и связности веб-графов, а именно:

1. Степень вершины (Degree Centrality) — показатель, указывающий для каждой страницы количество страниц, связанных с ней. В веб-графе вычисляются полустепень захода (indegree, количество входящих) и полустепень исхода (outdegree, количество исходящих) ссылок [9, 10].
2. Мера центральности Betweenness — показатель, указывающий насколько часто данная веб-страница лежит на кратчайшем пути между всеми парами страниц сайта [9, 10].
3. Мера центральности Closeness — среднее расстояние от данной веб-страницы до всех остальных страниц сайта [9].
4. Мера центральности PageRank — показатель важности страницы. Чем выше ее показатель, тем она важнее.
5. Мера связности p-Cliques — подграф, который представляет собой полный граф, где p — количество вершин в данном подграфе [10].
6. Мера связности k-Cores — подграф, в котором каждая страница связана по крайней мере с k другими страницами в этом подграфе [10].

А также использовались такие общие показатели топологии, как:

1. Расстояние (Distance) — средняя длина всех кратчайших путей в графе [6].
2. Диаметр (Diameter) — длина самого большого кратчайшего пути в графе [6].

В качестве сайтов университетского Веба были взяты следующие сайты из Мирового вебометрического рейтинга университетов [4]:

1. Сайт, занимающий первое место в общем рейтинге:
   - Сайт Гарвардского университета — ГУ (www.harvard.edu).
2. Сайты, занимающие первые места в рейтинге по Российской Федерации:

- Сайт Московского государственного университета — МГУ (www.msu.ru);
- Сайт Санкт-Петербургского государственного университета — СПбГУ (www.spbu.ru).

Используя вышеуказанный программный комплекс, требовалось получить визуальное представление топологии сайтов университетов в виде композиции наилучших значимых веб-страниц по всем мерам центральности (по каждой мере выбирался топ-10 страниц с наилучшими весами), а также построить и оценить расстояние от главной страницы до этих страниц.

3. Результаты эксперимента. В ходе эксперимента для выбранных сайтов университетского Веба были получены следующие значения мер связности и общих показателей топологии (таблица 1).

Таблица 1. Теоретико-графовые характеристики внутренней топологии университетских сайтов

Теоретико-графовые характеристики	**ГУ**	**МГУ**	**СПбГУ**
Количество страниц	451	45 557	51 945
Количество ссылок, принадлежащих главному домену	27 615	2 043 221	4 862 750
Общее количество ссылок	72 355	2 143 474	5 017 815
Расстояние	3,04	7,19	6,02
Диаметр	7	31	19
Среднее значение indegree, outdegree	13,59	28,1	69,79
Количество вершин в подграфе p-Cliques	27	10	144

На рис. 1–3: ■ — главная страница сайта; ■ — полустепень исхода; ■ — полустепень захода; ■ — мера Betweenness; ■ — мера Closeness; □ — мера PageRank; значение слева — расстояние от главной страницы.

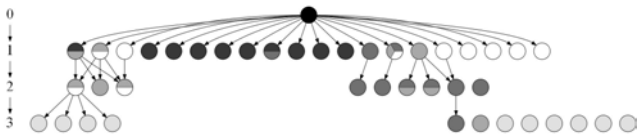

Рис. 1. Композиция всех наилучших значимых веб-страниц сайта ГУ

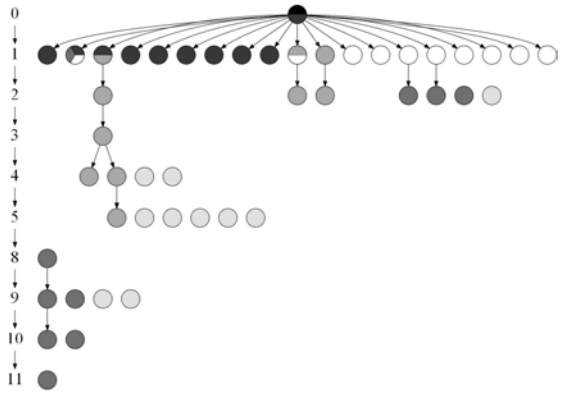

Рис. 2. Композиция всех наилучших значимых веб-страниц сайта МГУ

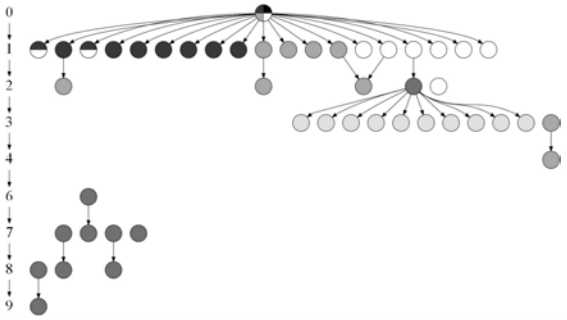

Рис. 3. Композиция всех наилучших значимых веб-страниц сайта СПбГУ

Композиция всех наилучших (по заданным мерам центральности) значимых веб-страниц исследуемых сайтов представлена на рис. 1–3.

4. Общие выводы. По результатам эксперимента можно сделать следующие выводы:

1. Сайт ГУ представляет собой навигационный сайт, перенеся большую часть информации на свои поддомены (ссылки главного домена составляют 38,2% от общего числа ссылок). Данное явление подтверждается тем, что веб-страницы по полустепени исхода у ГУ лежат близко к главной странице (глубина

1–3), в то время как у МГУ и СПбГУ они лежат в основном на определенном удалении (8–11 у МГУ и 6–9 у СПбГУ).

2. Показатели СПбГУ (такие как расстояние, диаметр, количество вершин в p-Cliques, положение значимых вершин относительно начальной страницы) лучше, чем таковые у МГУ. Это значит, что сайт СПбГУ показывает более высокую связность, нежели сайт МГУ.
3. Авторитетные страницы расположены на следующем уровне глубины после главной страницы, о чем говорит их высокий вес PageRank.
4. Значимые веб-страницы по мере Betweenness у всех сайтов лежат в радиусе пяти ссылок от главной страницы, однако у сайта МГУ они расположены последовательно. Это объясняется тем, что множество кратчайших путей проходит через эти страницы. При отказе одной из таких страниц увеличиваются размеры множества кратчайших путей, а также появляется риск полной потери связей с множеством страниц.
5. Значимые веб-страницы по мере Closeness у ГУ лежат на уровнях 2 и 3, у МГУ — в основном на уровнях 4 и 5, у СПбГУ — на уровне 3. Это показывает, что рассмотренные сайты централизованы не далеко от главной страницы.

Также был получен интересный результат: главная страница ГУ не попала ни в один из топ-10; главная страница МГУ — лишь в топ-10 по степени захода; главная страница СПбГУ — в топ-10 по степени захода, по мере Betweenness и по мере PageRank.

В дальнейшем планируется расширить эксперимент, проверив эргономические параметры значимых веб-страниц, выделенных программным комплексом, у заданных сайтов.

Литература

1. Almind T. C., Ingwersen P. Informetric analyses on the World Wide Web: Methodological approaches to «webometrics» // Journal of Documentation. 1997. No 53 (4). P. 404–426.

2. Thelwall M. Webometrics and Social Web Research Methods. University of Wolverhampton. [Электронный ресурс]: URL:http://www.scit.wlv.ac.uk/~cm1993/papers/

IntroductionToWebometricsAndSocialWebAnalysis.pdf (дата обращения: 2.03.15).

3. Печников А. А. Вебометрические исследования Web-сайтов университетов России // Информационные технологии. 2008. № 11. С. 74–78.

4. Ranking Web of Universities (Main page). [Электронный ресурс]: URL:http://www.webometrics.info (дата обращения: 2.03.15).

5. Блеканов И. С., Сергеев С. Л., Максимов А. Ю. Веб-краулер как инструмент для вебометрических исследований на примере анализа Веб-пространства СПбГУ // Materialy IX mezinarodni vedecko-prakticka konference «Moderni vymozenosti vedy – 2013». Praha, Czech Republic: Publishing House «Education and Science» s.r.o, 2013. Vol. 71. P. 66–69.

6. Блеканов И. С., Максимов А. Ю. Вебометрические исследования сегмента университетского Веба с помощью поискового робота // Процессы управления и устойчивость: Труды 44-й международной научной конференции аспирантов и студентов / под ред. Н. В. Смирнова, Т. Е. Смирновой. СПб.: Издат. Дом С.-Петерб. гос. ун-та, 2013. С. 403–408.

7. Blekanov I. S., Sergeev S. L., Maksimov A. I. Analysis of the topology of large Web segments using Broder's bow-tie model // Life Science Journal. 2014. No 11. P. 258–261.

8. Блеканов И. С., Сергеев С. Л., Мартыненко И. А. Построение тематико-ориентированных веб-краулеров с использованием обобщенного ядра // Научно-технические ведомости СПбГПУ. 2012. № 5 (157). С. 9–15.

9. Wasserman S., Faust K. Social Network Analysis. Methods and Applications. Cambridge: Cambridge University Press, 1994. 857 p.

10. Ortega J. L., Aguillo I. F. Visualization of the Nordic academic web: Link analysis using social network tools // Information Processing and Management: an International Journal. 2008. No 44 (4). P. 1624–1633.